

# Object Identity and Life Science Research

Robert J Robbins

Fred Hutchinson Cancer Research Center

[rrobbins@fhcrc.org](mailto:rrobbins@fhcrc.org)

# **SEMANTIC WEB: ISSUES**

# Object Identity and Life Science Research: Issues for the Semantic Web

- ▶ In any semantic web for the life sciences, no matter what technology is used, several needs must be met:
  - **IDENTITY MANAGEMENT:** It must be possible to identify unambiguously biological objects (more precisely to identify digital objects and associate them unambiguously with real-world biological objects).
  - **IDENTITY ADJUDICATION:** It must be possible to determine whether two different digital objects describe the same or different real world objects
  - **REFERENTIAL INTEGRITY:** It must be possible to make unambiguous, semantically well-defined assertions linking an object in one information resource to one or more objects in other information resources.

# Object Identity and Life Science Research: Issues for the Semantic Web

- ▶ In any semantic web for the life sciences, no matter what technology is used, several needs must be met:
  - RETAIL VS WHOLESALE CUSTOMERS: The semantic web must support the retail needs for coherence and the wholesale need for variation and disagreement (cf elephant and blind men story)
  - TRI\_STATE LOGIC: Systems involving the classification of biological objects need tri-state logic to handle queries.
  - NO CURATION: In all but the best-funded public databases, there are no funded resources available for information curation.
  - CONSISTENCY IS IMPOSSIBLE: science consists of assertions and observations, not facts; assertions and observations can differ without being untrue.

# Object Identity and Life Science Research: Issues for the Semantic Web

- ▶ In any semantic web for the life sciences, no matter what technology is used, several needs must be met:
  - FINAL ONTOLOGY REQUIRES PERFECT KNOWLEDGE: In a context-free global environment, the data model must meet the requirements of all possible users (or fail for some users).
  - REALITY IS NOT NEGOTIABLE: The requirements for scientific information systems are determined by discovery, not negotiation.
  - SOCIOLOGICAL IMPEDIMENTS: Technological solutions must also meet sociological requirements; an information system that could manage useful information is a failure if many are unwilling to participate.
  - EXPECTATIONS MUST BE MANAGED: never forget,

$\text{success} = \text{deliverables} / \text{expectations}$

# BACKGROUND ISSUES

# Philosophical Background: Identity

- ▶ Concept of identity still subject to metaphysical distinctions:
  - NUMERICAL IDENTITY: one thing being the one and only such thing in the universe - e.g., there should be one and only human being associated with a patient ID
  - QUALITATIVE IDENTITY: two things being identical (sufficiently similar) in enough properties to be perfectly interchangeable (for some purpose) – e.g., there are many books associated with an ISBN identifier

# Philosophical Background: Properties

- ▶ Properties are subject to identity-related distinctions:
  - ACCIDENTAL PROPERTIES: properties of an object that are contingent – that is, properties that are free to change without affecting the identity of the object
  - ESSENTIAL PROPERTIES: non-contingent properties – that is, properties which DEFINE the identity of the object and thus which cannot change without affecting the identity of the object (for some purpose)



## Philosophical Background: Properties

- ▶ Properties are subject to identity-related distinctions:

**Recognizing the distinction between essential and accidental properties will be critical in developing a successful identifier scheme for caBIG.**

**Especially challenging will be the fact that whether a particular property is essential or not is often context dependent.**

s,  
ich  
ome

# Philosophical Background: Properties

- ▶ Properties are subject to identity-related distinctions:
  - INTRINSIC PROPERTIES: properties of an object that are properties of the thing itself
  - EXTRINSIC PROPERTIES: properties of the object that are properties of the object's relationship to other objects external to itself

## Philosophical Background: Properties

- ▶ Properties are subject to identity-related distinctions:
  - INTRINSIC PROPERTIES: properties of an object that are properties of the thing itself
  - EXTRINSIC PROPERTIES: properties of the object that are properties of the object's relationship to other objects external to itself

**Identifying tandemly duplicated genes is a perfect example of the need to distinguish between extrinsic and intrinsic properties.**

## Philosophical Background: Identification

- ▶ “Identification” is a process that reduces ambiguity. Ambiguity reducing identification can occur in a number of different ways:
  - INDIVIDUAL SPECIFICATION: denoting an individual object without identifying either its class membership or its individuality - e.g., “this thing”
  - CLASS IDENTIFICATION: specifying that an object is a member of a class of objects that are sufficiently similar that the objects may be considered interchangeable (for some purpose) – e.g., “this book is Darwin’s *Origin of Species*”
  - INDIVIDUAL IDENTIFICATION: specifying that an object is in fact a PARTICULAR genuinely unique object in the universe – e.g., this book is Darwin’s own personally annotated copy of *Origin of Species*”

## Philosophical Background: Identification

- ▶ “Identification” is a process that reduces ambiguity. Ambiguity reducing identification can occur in a number of different ways:

**Note that as we move along this continuum our notion of “essential properties” changes.**

**This shows again that the concept of identity can be context dependent.**

- INDIVIDUAL IDENTIFICATION: specifying that an object is in fact a PARTICULAR genuinely unique object in the universe – e.g., this book is Darwin’s own personally annotated copy of *Origin of Species*”

## Practical Issues: Identifying What?

- ▶ Digital identifiers (IDs) perform different kinds of identification:
  - REAL-WORLD IDENTIFIER: identifier serves as a digital token representing a real-world (i.e., non-digital) object (e.g., patient ID); this kind of identifier is often used to associated a digital object (bag of properties) with a real-world object
  - DIGITAL IDENTIFIER: identifier serves as a digital token representing a (published?) digital object (e.g., LSID or URL)

## Practical Issues: Identifying What?

- ▶ Digital identifiers (IDs) perform different kinds of identification:
  - REAL-WORLD IDENTIFIER: identifier serves as a digital token representing a real-world (i.e., non-digital) object (e.g., patient ID); this kind of identifier is often used to associated a digital object (bag of properties) with a real-world object
  - DIGITAL IDENTIFIER: identifier serves as a digital token representing a (published?) digital object (e.g., LSID or URL)

**This distinction can be hard to make:  
What does an IP address identify?**

## Practical Issues: Identification vs Specification

- ▶ Digital identifiers (IDs) can truly identify particular objects or they can merely specify singular objects, with no guarantee of what that singular object is:
  - IDENTIFICATION: the same LSID should always return exactly the same (bit for bit) digital object
  - SPECIFICATION: the same URL is not guaranteed to return the same thing twice



## Practical Issues: Identification vs Specification

**Note that these two situations really just represent the opposite ends of a continuum:**

**At one end EVERY property is essential – at the other end NO property is essential.**

**At both ends, the relationship of identifier to object is clear. In between, this clarity does not exist and contention can and will exist between identifiers and properties (e.g., the same human being could accidentally be assigned two patient IDs, but we could infer identity from the essential properties).**

## Practical Issues: Identity Claims

- ▶ Different methods exist for answering the question whether or not two objects are the same :
  - DEMONSTRATED IDENTITY: the identifiers are the same and the essential properties are the same
  - INFERRED IDENTITY: the identifiers are different but the essential properties are the same
  - INFERRED NON-IDENTITY: the identifiers are the same, but the essential properties are different
  - ASSERTED IDENTITY: the identifiers are the same, but the state of the essential properties are unknown

## Practical Issues: Identity Claims

- ▶ Different methods exist for answering the question whether or not two objects are the same :

**With checksums, LSIDs are an instance of DEMONSTRATED identity.**

**Without checksums, LSIDs are an instance of ASSERTED identity.**

- **ASSERTED IDENTITY**. the identifiers are the same, but the state of the essential properties are unknown